基于标签的商品推荐模型研究*

涂海丽 [1] 唐晓波 [2]
[1](东华理工大学经济与管理学院 南昌 330013)
[2](武汉大学信息管理学院 武汉 430072)

摘要: 【目的】构建社会化电子商务环境下基于标签的个性化商品推荐模型。【方法】综合考虑用户使用标签的频率和时间因素计算用户的兴趣偏好; 基于标签层次特征和电子商务网站中关于商品特征的检索条件, 构建某一主题商务社区中商品本体; 利用本体规范化用户标签语义, 并对商品进行分类; 寻找含有用户偏好的类簇, 计算该类簇中商品与用户偏好商品的相似度, 将用户未标注过的商品与用户偏好相似度高的商品推荐给用户。【结果】从翻东西网站上随机选取 200 个活跃用户关于热门商品的标注信息进行分析, 验证该模型的有效性。【局限】在计算用户兴趣偏好时, 只考虑用户使用标签的频率和时间因素, 未考虑其他因素。【结论】该模型相对于利用标签进行协同过滤推荐方法具有较优的效果, 计算时间和空间复杂度更小。

关键词: 用户标签 商品本体 用户偏好 推荐模型
分类号: G35

1 引 言

商品推荐的目标是综合运用各种方法建立用户兴趣偏好与商品之间的关联, 并主动呈现给用户。商品标签是用户关于商品描述的元数据, 本文研究的标签对象是社会化电子商务中用户自由标注的标签, 具有可挖掘的重要信息: 用户主动标注物品的行为反映了用户的认知模式和兴趣偏好; 标签能够反映物品特征。大量用户为物品添加描述性标签, 高频标签代表用户对相同物品特征的广泛认同; 标签具有可检索性。作为用户和物品间的桥梁, 标签系统一般提供通过标签检索物品的链接。社会化电子商务中由购物达人或普通用户自由标注的标签居多, 但会出现一词多义或一义多词的现象, 使标签的词表变得庞大。由于标签的大众化特征, 同一社区的很多标签都是杂乱无章的, 标签与用户、标签与物品之间可以是多对多的关系, 加大标签组织和利用的难度, 使得标签相似度计算不准确。因此, 作为一种原生态的自然语言, 标签语义的模糊性(即一词多义)、标签形式的多样性(即一义多词)和标签结构的扁平化(缺乏直接的层次逻辑关系), 极大地限制了其在个性化推荐中的作用, 在基于标签的推荐系统中, 推荐准确性低, 用户体验差。如何减少标签冗余和歧义给推荐带来的干扰、在扁平化的标签列表中发现它们之间的关联, 从而明确标签所表达的语义和主题, 是更好地将标签应用于商品推荐的关键。本文主要讨论社会化电子商务中 UGC 标签的应用, 研究如何利用本体序化用户标签及商品标签, 从中获取用户偏好及商品特征的主题描述, 探讨如何建立用户偏好与商品特征之间的关联, 从而为用户推荐个性化商品。

2 相关研究

根据推荐算法的不同, 国内外对基于标签的推荐研究方法主要归纳为以下几种。

通讯作者: 涂海丽, ORCID: 0000-0001-8325-9840, E-mail: 69417380@qq.com。

*本文系国家自然科学基金项目“社会化媒体集成检索与语义分析方法研究”(项目编号: 71273194)、抚州市社科规划项目“基于KANO 模型的抚州旅游市场需求分析”(项目编号: 15sk23)和东华理工大学地质资源经济与管理研究中心开放基金项目“基于 KANO 模型的旅游用户潜在需求挖掘研究”(项目编号: 14GL05)的研究成果之一。

(1) 矩阵分解。将用户、用户标注的资源以及标注的标签三者之间的三元关系矩阵分解成两两组合的二维矩阵，先发现两两之间的关系，再进行综合，找到三者的对应关系，这样既可以减少矩阵计算复杂性，也能够实现标签或资源的推荐[1-3]，该方法是基于标签的推荐系统中的研究热点之一。

(2) 张量分解。该方法不进行三元矩阵分解，而是利用奇异值分解的方法进行降维，然后排序标签，实现标签推荐；也可以根据标签与资源的关联关系，向用户推荐资源。

(3) 聚类方法。将标签、用户和资源分别进行聚类，具体如下：

①用户聚类。根据现有标签或资源的相似可以推测用户兴趣的相似，相似用户会有更多潜在共性，甚至可以结成一个特殊的社群[4-5]。

②资源聚类。通过资源聚类发现资源中的"睡美人"，提高资源推荐的覆盖率。

③标签聚类。这是当前标签推荐或利用标签的资源推荐需要聚类时的首选。主要是依据标签共现次数聚类标签，常用的聚类算法有 K-means、Markov 等，利用聚类结果中标签之间的关联，计算对应资源间的相似度，进行资源推荐[6]。Niwa 等[7]在利用 TF-IDF 公式计算标签权重的基础上聚类标签，据此计算用户偏好资源与聚类中标签对应资源的相似度，实现资源推荐。Gemmell 等[8]对标签进行层次聚类，基于此构建用户兴趣模型。杨丹等[9]通过标签聚类计算用户与标签的相似度，实现网页推荐。

(4) 图论方法。该方法利用网络图表达用户、用户标注的资源以及标注的标签三者之间的关系，利用社会网络分析方法进行用户偏好建模，从而实现基于内容的资源推荐或资源协同推荐[10-11]。图论方法中的典型代表是 Hotho 等[12]研究出的 FolkRank 算法。该算法利用无向图表达用户、用户标注的资源以及标注的标签三者之间的关系。图中的节点是三者的并集，边是两两之间的共现值，通过对图中各元素的关联度分析，找出重要标签并排序，将重要标签对应的资源推荐给用户。构图只是基础，重要的是对图的分析，社会网络分析方法才是图论方法的核心，受到学者们的重点关注。

此外，还有一些其他的研究视角和方法。如 Schmitz 等[13]利用数据挖掘技术中的关联规则挖掘研究对象的分类结构特征，进行人员、标签和项目的推荐。曹高辉等[14]认为每个标签可以看成一个概念，标签集合构成概念空间，并具有层次结构，通过构建标签层次结构实现资源的个性化推荐。田莹颖[15]认为用户标注行为存在兴趣漂移的问题，提出利用 TF-IDF 和后控词表，给用户最近标注的标签设置较高的时间权重，计算用户之间的相似度，找出共同标注的信息资源，并通过标签对用户与资源进行匹配，将相匹配的信息推荐给目标用户。邓双义[16]将标签作为媒介，利用 WordNet 语义，计算用户偏好的标签集与资源的标签集的相似度，将相似度高的标签分别对应的用户和资源进行比对，并将相匹配的资源推荐给用户。还有将以上主要方法相结合的混合推荐方法，Rafailidis 等[17]先对标签、用户和资源三阶矩阵利用张量分解方法降维，然后对标签聚类，既解决了三元矩阵计算复杂度高的问题，也避免了稀疏矩阵对相似度计算的影响，对两种方法扬长避短，实现了资源的个性化推荐。还有根据标签的流行度、时间特征或标签的代表性、用户与标签的亲和力等刻画用户对资源的偏好，采用梯度下降法对用户–资源矩阵进行分解，利用分解后的特征矩阵对目标用户进行预测并推荐[18-19]。

虽然学者们从多个视角研究了基于标签的推荐算法来解决推荐研究中固有的问题，并试图避免标签本身的缺陷带来的新问题。但是，这些基于标签的推荐算法仍然存在如下不足：

(1) 不管是矩阵方法还是图论方法，计算复杂度都很高；

(2) 虽然标签总量较大，但部分单个用户所标注标签数目较少，难以准确获取用户偏好，限制了推荐的效果；

(3) 标签语义存在歧义，造成数据的噪音干扰；

(4) 目前大部分研究在进行推荐时假设用户兴趣是不变的，这不符合现实情况，虽然最近有些研究考虑了时间等情境因素对用户标注行为的影响，但很少考虑多方面因素的综合影响，且研究成果较少。

因此，在前人研究基础上，本文提出一种社会化电子商务环境下利用社会化标签的个性化商品推荐模型，该模型综合考虑用户使用标签的频率和时间因素计算用户的兴趣偏好，并基于标签特征和电子商务网站中商品检索条件，构建某一主题商务社区中商品本体，利用本体规范化用户标签语义，并对商品进行分类，寻找含有用户偏好的类簇，计算该类簇中商品与

用户偏好商品的相似度，并将用户未标注过的与用户偏好相似的商品推荐给用户。本文方法旨在对算法计算的复杂度、标签语义规范化、以及综合考虑不同因素对标签作用的影响三方面进行改进。

3　基于标签的商品推荐模型构建

根据以上的思路，构建基于标签的商品推荐模型，如图 1 所示。

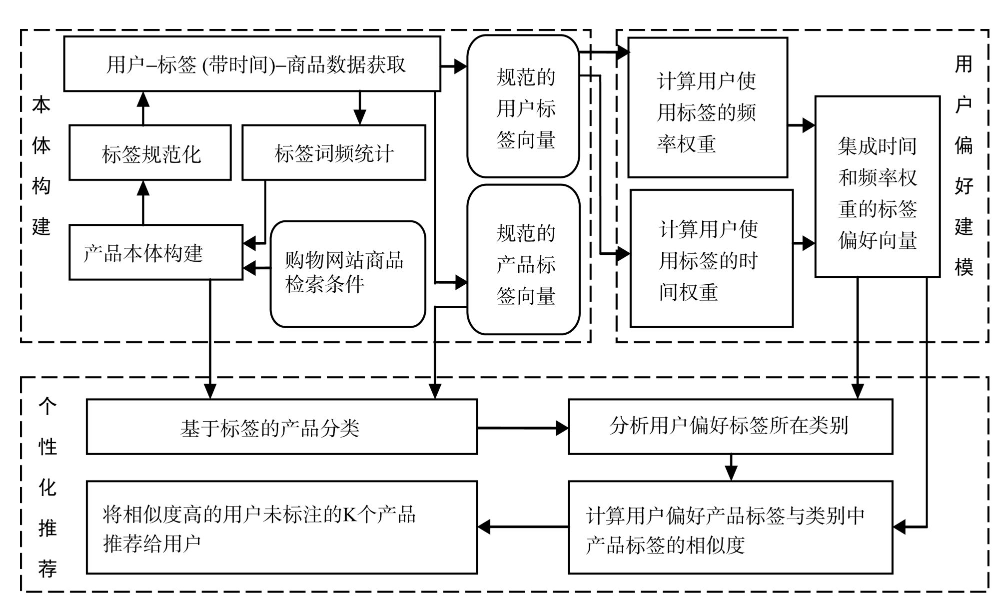


图 1　基于标签的商品推荐模型

3.1　商品本体构建

(1) 标签数据获取及词频统计

在社会化电子商务网站中，每一个注册用户可以自由管理感兴趣的商品信息。很多社会化电子商务网站提供了用户分类表达自己兴趣内容的工具：如“喜欢”、“兴趣”、“关注”、“分享”等分类夹。翻东西网让用户将自己满意的试穿效果图放在“哇晒”分类夹中，而将自己在其他购物网站看到并喜欢的商品通过复制网址的方式分享于“喜欢”分类夹中，在“帮我挑”中分享自己的购物经验，用户也可以关注其他用户或品牌。由于用户标注数量相差较大，大部分用户标签稀疏，本文对商务社区中标签的理解不单是“喜欢”分类中的标签，而是所有分类夹中用户对商品的标注，以全面获取用户兴趣偏好。

购物社区中的用户标签不仅是用户利用简短关键词对商品名称和商品特征的个性表达，也是用户与商品之间的纽带。通过观察不同用户对同一商品的标注关系以及一个用户对多个商品的标注关系组成的集合，可以看出用户、标签和资源三者之间的关联。这样可以通过合适的方法，将标签作为中介和分析对象，发现用户关于商品的兴趣偏好，如图 2 所示。

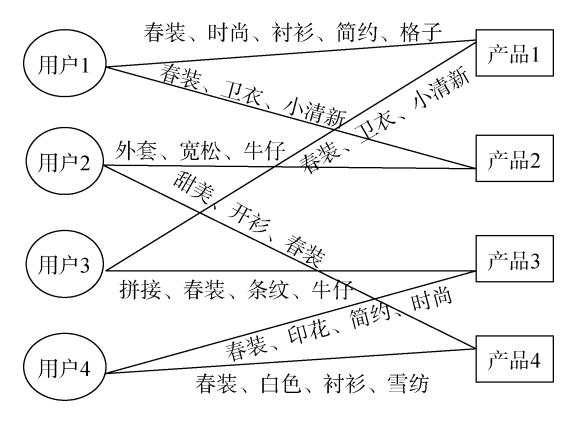


图 2　社会化电子商务中的用户–标签–商品关系示例

社会化电子商务网站提供了用户给商品添加标签的功能，并通过积分奖励的办法鼓励他们将自己喜欢的商品和标注的标签分享给网站中其他的注册用户，当然这些分享信息非注册用户也可以看到。这些标签都在用户的“喜欢”、“晒单”、“兴趣”、“分享”主题下，显示了用户感兴趣的商品及偏好主题，那么形式化表达用户–标签–商品之间的关系是用户偏好获取的前提。

本文先利用网络爬取工具从商务社区爬取用户、用户标注的商品标签及其时间(各标签之间用空格分开)、

该用户标注过的商品信息,并将其保存在电子文档中。将标签分别表示为用户标签(某用户标注的所有商品标签)和商品标签(不同用户给同一个商品的标注),用户标签初始表示为 $i((tag_1, time_1), (tag_2, time_2), \cdots, (tag_n, time_n))$, i 为用户集合 I 中的元素, n 为用户 i 所使用的标签数, $time$ 为对应标签标注的时间。商品标签初始表示为 $p((tag_1, freq_1), (tag_2, freq_2), \cdots, (tag_m, freq_m))$, p 为商品集合 P 中的元素, m 为商品 p 所使用的非重复标签数, $freq$ 为对应标签使用的次数。将电子文档中标签一列单独取出,利用中国科学院计算技术研究所的ICTCLAS3.0分词系统对其进行词频统计,并按词频大小排序标签。

(2) 商品本体构建方法

本体能够表达概念之间的语义层次关系,利用标签本体可以规范标签语义,也可以进行标签分类。本文构建标签本体的目的是对商品类型和商品属性等信息进行规范化的再组织,以提高商品推荐的效果。遗憾的是,由于本体构建本身的难度,到目前为止,很少有将本体应用于基于标签的商品推荐中的研究成果,说明基于标签的推荐与本体相结合的研究还很少见。

标签不仅能够表达用户偏好,也标注了商品属性和类别,隐含表达了各种商品及其属性的层次关系和对应关系。本文实验研究的是服饰类商品的标签本体和推荐问题,图3表示商品和标签的层次及对应结构示意图。其表达的意思是,一个大类下面有多个小的类别,如服饰与帽子、上衣;每一个小类可以有多个实例,如裤子小类中包含打底裤、短裤等;每一个实例可以有多个特征,如用户可以对一条长裙标注清新、优雅等多个标签,这些大类、小类、实例和特征之间具有层次关系,标注它们的标签也应该具有层次关系。

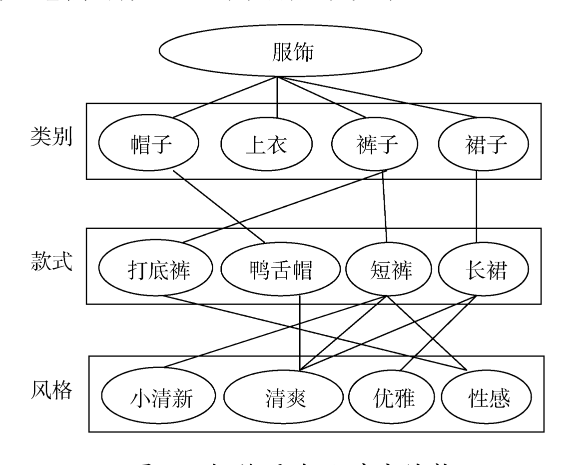


图 3 标签层次及对应结构

一个商品可以用多个标签标注,而标签之间的层次及对应关系可以表达出来。那么,在基于标签的商品推荐中,通过这种层次结构对商品分类,并基于标签计算商品之间的相似度时,具有以下规律:

①描述不同商品特征共用的标签越多,而共用标签标注的商品越少,这些商品越相似。图3中,"打底裤"和"短裤"两款商品共用的标签是"性感",而"鸭舌帽"、"短裤"、"长裙"三款商品共用的标签是"清爽",因此 sim(打底裤, 短裤)>sim(长裙, 短裤);

②共同标签离商品越远,商品之间越不相似,反之越相似。图3中商品"打底裤"、"短裤"的最近共同标签是"裤子",商品"短裤"、"长裙"的最近共同标签是"服饰",而"裤子"是"服饰"的子节点,因此 sim(打底裤, 短裤)>sim(长裙, 短裤);

③由于标签是对商品的全方位描述,理论上,商品的标签差异性越大,商品越不相似,反之越相似。但由于有些标签会重复使用,实际的差异性可能比按相似度计算出来的更大。

社会化电子商务网站中的用户标注的标签随意性很大,也很难看出其层次对应关系,因此,本文参照电子商务购物网站淘宝网的服饰类搜索条件,构建服饰类商品标签本体的品种及其属性关系。如输入"服饰",其下品类有"衣服"、"鞋子"、"首饰"、"包包"等;"衣服"品类下有"上衣"、"裙子"、"裤子"等,而关于"裙子"特征的检索条件又有"材质"、"图案"、"风格"、"流行元素"。再根据社会化电子商务网站翻东西中"大家淘"版块的热点标签中用户给服饰类商品标注的标签的词频统计,构建商品本体。如在裙子的"流行元素"特征描述中,高频词有"拼接"、"镂空"等,"风格"特征描述的高频词有"时尚"、"可爱"等。结合淘宝网和翻东西网上的热点标签构建的商品标签概念本体,如图4所示。

(3) 标签的规范化处理

根据本体中不同商品类型及其属性描述词汇,特别是关于商品特征的描述词汇,将用户随意使用的属性词汇用本体中意思相同或最相近的属性描述词替换。这里参照电子商务网站的检索条件和社会化电子商务网站高频词,通过人工综合分析来替换。如以上的服饰商品中,对上衣风格的描述有"卡通"、"甜美"、"小清新"、"萌"、"可爱",其中"卡通"是本体中没有的,但根据同义词典,这些词都与"可爱"意思相近,因此用"可爱"替代。类似的还有"百搭"与"混搭",将"混搭"统一替换为"百搭",等等。将所有规范化的

标签更新电子文档中的初始标签，并将同一用户关于不同商品的标签和不同用户关于同一商品的标签分别表示为标签向量，作为下一步用户偏好建模的输入数据。

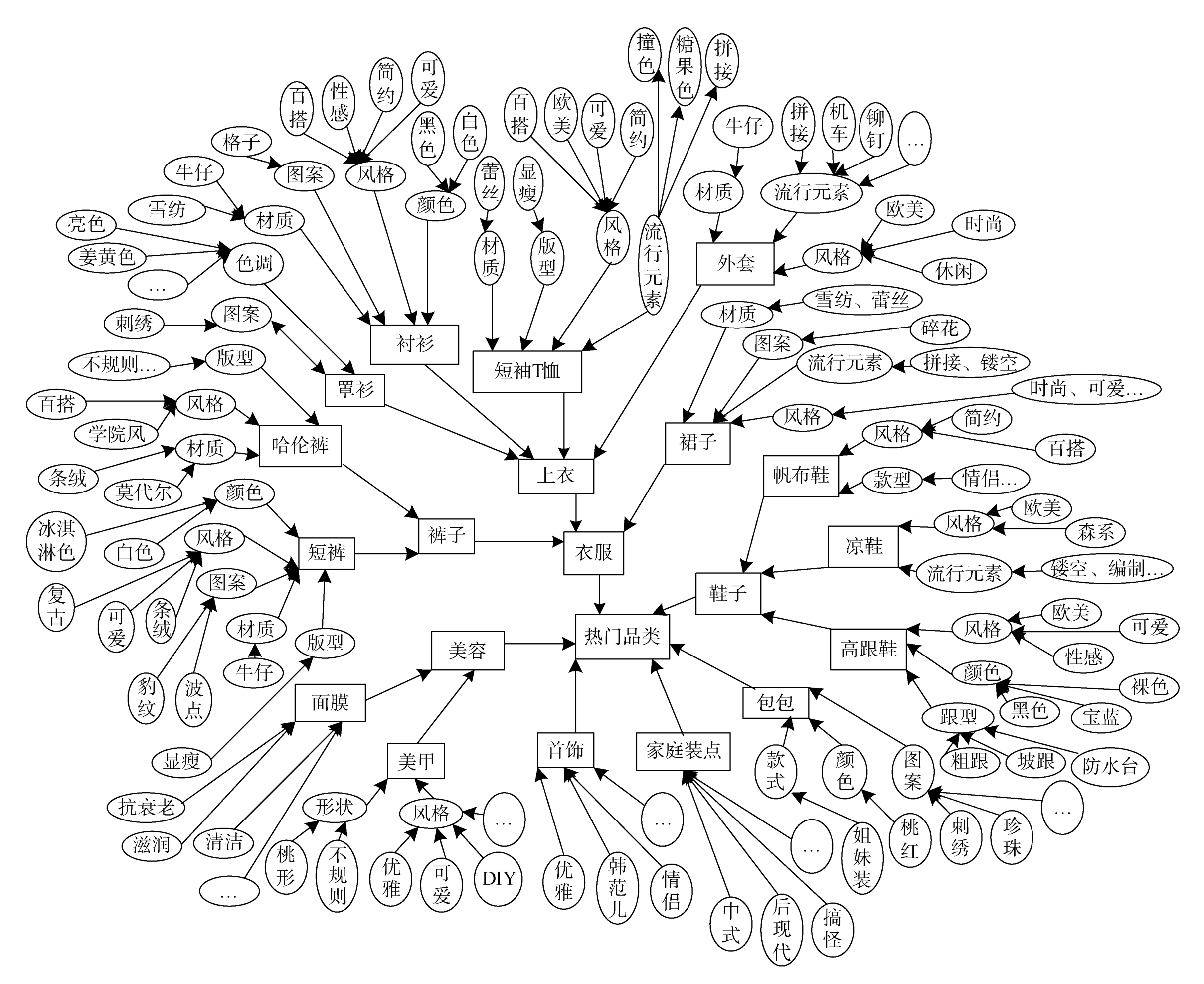


图4　基于标签的本体构建

3.2　用户偏好建模

根据标签构建用户偏好模型的目的是从标签中获取用户对商品的隐性需求或偏好。综合考虑标签标注时间和标签使用频率对用户偏好的影响，分别计算用户使用标签的时间权重和用户使用标签的频率权重，集成这两个影响因子权重计算基于标签的用户偏好。

(1) 用户使用标签的频率权重计算

用户使用的商品标签能够反映该用户对商品的兴趣偏好。用户对某些标签使用越多，说明对其情有独钟，也说明对这些标签共同描述的商品的喜爱。对于商品而言，不管多少个用户对一件商品进行标注，标注的标签可以反映该商品的特征。某些标签使用的频率越高，它们就越能代表这个商品的特征。

在构建用户偏好模型时，重点考虑用户使用过的标签。当用户在浏览网上资源时，对自己喜欢的资源选择相应的标签进行标注。标签的类别可以从一定程度上反映用户的喜好类型，比如用户“好男人”采用的标签中，经常出现“外套”、“休闲”等短语，那么“好男人”可能喜欢外套或休闲类服饰。并且“好男人”使用的标签中“休闲”这一标签出现的频率较高时，可能因为用户更加喜欢休闲类服装。也就是说，用户使用的标签频率可以反映用户的喜好程度。但有学者提出，如果用户高频率使用标注系统中低频率出现的标签，则表明该标签内容更能反映用户对商品的偏好。故此通过计算标签与用户间的关联程度可以判断标签内容是否真正与用户兴趣相吻合。这也是传统TF-IDF算法的要义，进而

引入 TF-IDF 算法可以计算标签与用户的关联程度。

假设 U 表示用户集合，T 表示标签集合，P 表示商品集合。

①对 $u\in U$，P_u 表示用户 u 标注过的商品集合，T_u 表示用户 u 使用过的标签集合。

②对 $t\in T$，P_t 表示用标签 t 标注过的所有商品集合，U_t 表示用过标签 t 的用户集合。

③对 $p\in P$，T_p 表示标注了商品 p 的标签集合，U_p 表示标注了商品 p 的用户集合。

因此，三元组(u, p, t)表示用户 u 用标签 t 标注了商品 p。

每个用户的标签集通过使用一个标签向量 $T_u=(t_1u(f_1), t_2u(f_2), \cdots, t_mu(f_m))$来表示。其中，$m$ 是标签的个数，t_mu 表示用户 u 的第 m 个标签，f_m 表示用户 u 的第 m 个标签的频率。$t_mu(f_m)$描述标签 t_m 表示的用户偏好程度，f_m 用 TF-IDF 公式计算，如公式(1)所示。

$$t_m u(f_m) = Tf_u(f_m) \times IDF_u(f_m) \tag{1}$$

对三元组进一步挖掘，将用户 u 使用的标签 t 的次数记为 $count(u, t)$，使用标签 t 标注商品 p 的用户集合记为 $UserCount(t, p)$。由用户使用的标签 t 的次数及所有用户使用标签 t 的次数，可以得出用户 u 使用标签 t 的频率，用 $Tf_u(f_t)$表示该频率，如公式(2)所示。

$$Tf_u(f_t) = \frac{UserCount(u,t)}{\sum_{k\in U_t} UserCount(u,k)} \tag{2}$$

其中，k 表示用户标注过的某个标签，如公式(3)所示。

$$IDFu(f_t) = \log\frac{N}{n_t} \tag{3}$$

其中，N 表示用户总数，n_t 表示收藏和使用标签 t 的用户总数。

将公式(2)和公式(3)带入公式(1)得出用户与标签的联系程度，如公式(4)所示。

$$t_m u(f_m) = \frac{UserCount(u,t)}{\sum_{k\in U_t} UserCount(u,k)} \times \log\frac{N}{n_t} \tag{4}$$

商品的各个标签的使用频率也可以用标注该商品某标签的使用次数除以该商品的所有标签数。那么标签与商品的相关程度的计算方法如公式(5)所示。

$$relate(t,p) = \frac{UserCount(t,p)}{\sum_{q\in P_t} UserCount(t,q)} \tag{5}$$

其中，q 表示由标签 t 标注的某个产品，$relate(t, p)$值越大，说明表示用该标签标注的同一产品的用户越多，该标签与产品的相关度越大，标签 t 越能代表产品 p，在标签 t 下产品 p 得到推荐的优先级就越高。这样，产品的标签向量可以表示为 $T_p=(t_1p(relate(t_1, p)), t_2p(relate(t_2, p)), \cdots, t_mp(relate(t_m, p)))$。

(2) 用户标注标签的时间权重计算

由于用户兴趣存在偏移现象，用户所使用的标签会随时间而变化。例如，用户 u 过去用众多“春装”的标签去标注相关商品，也许因为那时是春秋季节，此人想购买当季的服装。随着季节变化，用户可能会关注其他季节的服装。再如，当用户计划旅游时，会关注旅游地以及旅游景点进而对这些信息有较多的标注，当用户选择一个旅游景点后，可能会关注当地的宾馆、小吃、特产以及娱乐场所等。因此应更关注用户近期的标注，这种近期标签相比历史标签更能反映用户兴趣热点，对用户未来行为预测更有帮助。所以，时间是标注行为中的重要信息因素，引入时间信息能更好地获取用户最新兴趣热点，使用户获得高匹配的个性化推荐。

通常，用户关注资源的时间距当前越近，该资源就越有价值，即与用户当前兴趣热点相关性越高。另外，用户对标签的兴趣偏好与用户对同一标签关注时间长度正相关，关注时间持续越长，用户对标签越感兴趣，标签与用户当前兴趣热点吻合度越高。Cheng 等[20]考虑到用户兴趣热点会随时间偏移，采用自适应指数衰减函数来处理这一问题，而指数遗忘函数是利用时间效应建模中广泛使用的一种函数，这种方法通过弱化用户历史行为影响以强化近期行为的作用。本文将指数遗忘函数引用到通过用户对标签使用时间来挖掘用户标签偏好中，结合时间信息计算用户标注的标签权重，如公式(6)所示。

$$P_{time}(u_m, p_n) = \exp\{-\ln 2 \times time(u_m, p_n) / hl_u\} \tag{6}$$

其中，$P_{time}(u_m, p_n)$是通过时间因素计算出来的用户 u_m 对产品 p_n 的标签权重，揭示了用户 u_m 对产品 p_n 的偏好。其中 $time(u_m, p_n)$是一个非负整数值，当用户 u_m 对产品 p_n 的标注行为是用户 u_m 的标注行为的最后一天，那么 $time(u_m, p_n)$被设置成 0，若是倒数第二天，则设置为 1，以此类推。hl_u 代表用户的生命周期，其计算方法如公式(7)所示。

$$hl_u = Date_{last} - Date_{begin} \quad (7)$$

其中，$Date_{last}$ 是用户最后一次标注标签的时间，$Date_{begin}$ 是用户第一次标注标签的时间。

长生命周期用户的兴趣因稳定性好而下降缓慢，故对其近期兴趣不宜过高偏重。而短生命周期用户兴趣因不成熟性而变化较快，生命周期短的用户的兴趣变化大，故对其近期兴趣应给更多倚重。本文赋予近期行为权重高于之前历史行为权重，借助时间效应更好地识别出用户当前兴趣热点。

使用时间权重对每个用户的标签集进行量化表示，通过使用一个标签向量 $T_u=(t_1u(time_1), t_2u(time_2), \cdots, t_mu(time_m))$表示。其中，$m$ 是标签的个数，t_mu 表示用户 u 的第 m 个标签，$time_m$ 表示用户 u 的第 m 个标签的时间权重，$t_mu(time_m)$描述标签 t_m 在多大程度上体现近期用户 u 的兴趣爱好。

(3) 集成频率与时间的用户偏好表达

加权标签能更好地将用户对商品的意见与兴趣表现出来，其丰富信息有助于构建更全面和更精确的用户模型。用户对标签内容偏好程度与用户使用标签的频率正相关，用户越频繁使用某些标签，说明用户越偏爱这些标签所标注的商品；用户当前兴趣与标签使用时间负相关，即标签使用时间距当前时间越远，越不能反映用户当前兴趣，最新使用的标签则更能反映用户的当前兴趣。该模型利用上述两点提出频率权标签偏好和时间权标签偏好，最后将两者融合提出最终的用户标签偏好向量，使个性化推荐系统具有更好的可扩展性和实时性特征。

本文不仅利用用户对标签使用次数的多少评判其标签偏好，也考虑用户标注时间因素，即二者的集成。如果用户高频率使用某标签，说明此标签所标注的商品对用户具有高兴趣度，而标注的时间权重越大，越能够反映用户的最近兴趣。因此，本文用标签的频率权重与时间权重相乘得到用户最终的兴趣标签。利用上述公式对用户 u 的标签向量进行修正后的结果如公式(8)所示。

$$T_u=(t_1u(f_1)\times t_1u(time_1), t_2u(f_2)\times t_2u(time_2), \cdots, t_mu(f_m)\times t_mu(time_m)) \quad (8)$$

3.3 个性化商品推荐

本文提出的基于标签–本体的商品推荐在建立商品标签本体和用户偏好模型后，利用本体中概念之间的关系分类商品标签，并将分类后的商品标签集与用户偏好标签匹配，找到相匹配的商品标签集后，进一步计算各匹配标签集中的商品标签与用户偏好标签的相似度，将相似度高的若干商品标签所标注的商品推荐给用户。

(1) 基于本体的商品分类

本体中实体之间的关联关系以及实体与其属性之间的层次关系可以直观显示，而商品标签也非直观表达了商品的类型、风格等信息，商品的标签和分类之间存在联系。可以将商品标签遍历本体的这些关系，将商品进行归类，每一个类代表一个主题。利用标签本体对商品进行归类后，每一个商品属于一个分类簇，将所有的商品分配到一个单独的类簇(即主题)，有若干个处在本体描述的关系结构树的不同位置的类簇。

本文提出基于标签的商品推荐是一个跨主题的推荐，对每一个用户标签，在各类簇中查找与其匹配的商品标签，至少有一个与用户标签相同。找到匹配的类簇后，计算该类簇中用户未标注的各商品标签与用户标签的相似度，相似度越高越优先推荐。

(2) 同类别的商品标签与用户标签的相似度计算

用户对某个物品打上标签，说明用户对此物品存在某种兴趣。用户对该标签内容兴趣越高，使用频率越大。利用标签为特定用户进行物品推荐计算时，先通过计算与目标用户的相似性挖掘出相似用户，再借助用户协同推荐算法为目标用户提供含有 N 个供选物品的推荐列表。本文中用户将获得与自身偏好相似度高的商品推荐，并计算商品标签用户偏好之间的相似度。

上文已将用户标签及商品标签进行了权重表达，用户标签向量如公式(9)所示。

$$T_u=(t_1u(f_1)\times t_1u(time_1), t_2u(f_2)\times t_2u(time_2), \cdots, t_iu(f_m)\times t_iu(time_i)) \quad (9)$$

商品标签向量如公式(10)所示。

$$T_p=(t_1p(relate(t_1, p)), t_2p(relate(t_2, p)), \cdots, t_jp(relate(t_j, p))) \quad (10)$$

先将全体物品标签与加权后用户标签采取向量表示，利用余弦相似度方法进行匹配主题内商品 p_j 的标签和用户 u_i 的标签之间的相似度计算，如公式(11)所示。

$$sim_{ij}(u_i, p_j) = \frac{\vec{u_i^*} \times \vec{p_j^*}}{\left\|\vec{u_i^*}\right\|\left\|\vec{p_j^*}\right\|} \quad (11)$$

设定预设阈值为 e，如果 $sim_{ij}(u_i, p_j)>e$，则商品与用户偏好相似。

(3) 相似度排序与商品推荐

根据计算的用户标签偏好向量与商品标签向量的相似度，当二者的相似性大于阈值时，则将该商品作为候选推荐对象，得出所有的候选商品后，按相似度值从大到小排序，最终选取 TOP-K 形成推荐商品列表，推荐给用户。

4　实验及结果分析

4.1　实验数据的描述与处理

“翻东西”是国内典型的第三方社会化电子商务网站，网站中聚集了大量的用户，他们以“标签+图片”的形式分享自己喜欢、感兴趣以及购买过商品信息，其他用户也可以关注自己或自己的标签，也可以评论用户分享的商品。目前，该网站聚集了大量用户，热门标注的商品接近 30 万件，热门标签约 75 万个。本文从热门标注的商品中随机择取近期最活跃的 200 个用户作为目标用户，主要涉及服饰类商品，标签总数超过 8 万。对用户及其标注的信息进行采集，获取的字段包括用户名、标签、标注的时间、商品名称，以电子文档保存。

按用户将数据集随机分成 10 份，选取 1 份作为测试集，另外 9 份组成训练集。按照本文构建的推荐模型的思路和方法进行实验。实验工具有：八爪鱼采集器、Protégé、Excel、ICTCLAS3.0。

4.2　数据分析

分别计算用户标签的使用频率权重系数和标注时间权重系数，并将其乘积作为用户标签权重，也即用户标签偏好值。图 5 是测试集中用户“裙子飞了”的标签权重计算结果。

用户名	标签	时间权重	频率权重	时间*频率权重
裙子飞了	连衣裙(23	0.672	0.996	0.669
	外套(239)	0.637	0.957	0.609
	风衣(76)	0.637	0.785	0.500
	复古(153)	0.637	0.748	0.477
	学院风(99	0.637	0.717	0.457
	简约(212)	0.573	0.732	0.419
	防晒(32)	1.000	0.288	0.288
	拼接(72)	0.672	0.389	0.262
	条纹(73)	0.573	0.446	0.256
	衬衫(91)	0.637	0.378	0.241
	卫衣(55)	0.528	0.452	0.239
	半身裙(36	0.637	0.372	0.237
	格子(55)	0.637	0.349	0.222
	甜美(83)	0.528	0.416	0.219
	宽松(79)	0.637	0.341	0.217
	T恤(90)	0.570	0.381	0.217
	毛衣(54)	0.573	0.372	0.213
	蕾丝(60)	0.672	0.316	0.212
	牛仔(62)	0.637	0.318	0.203
	荷叶边(42	0.528	0.304	0.160
	代购(44)	0.551	0.290	0.160
	可爱(35)	0.637	0.241	0.154
	开衫(37)	0.500	0.304	0.152
	露肩(29)	0.503	0.300	0.151
	波点(29)	0.573	0.261	0.150
	百搭(61)	0.573	0.228	0.131

图 5　用户标签权重计算结果(部分)

根据建立的商品本体，将训练集中商品按商品主题分类，找到用户标签匹配的主题。笔者通过计算测试集中用户标签偏好值与训练集中匹配主题下的商品标签的相似度，将相似度高(本文取相似度阈值为 0.5，即将相似度不小于 0.5 的商品作为候选推荐商品)的若干商品推荐给用户。

为了检验本文方法(TFT-Based)，将其与只考虑标签时间权重的推荐(TT-Based)、只考虑标签频率权重的推荐(FT-Based)和不考虑标签权重的推荐方法(T-Based)进行比较。参照利用标签进行资源推荐的两篇文章的评价方法[21-22]，以准确率(*Precision*)、召回率(*Recall*)和 *F-Measure* 值三个指标作为本文推荐方法结果的度量。准确率表示用户对所推荐商品感兴趣的概率，召回率表示用户感兴趣的商品被推荐的概率。两个概率值越高，表示该方法推荐的质量越好。对于用户 u，令 $P(u)$为给用户 u 的长度为 N 的推荐列表，令 $D(u)$是测试集中用户 u 实际打过标签的物品集合。

计算如公式(12)–公式(14)所示。

$$Precision=\frac{\sum_{u\in U}|P(u)\cap D(u)|}{\sum_{u\in U}|P(u)|} \tag{12}$$

$$Recall=\frac{\sum_{u\in U}|P(u)\cap D(u)|}{\sum_{u\in U}|D(u)|} \tag{13}$$

$$F\text{-}Measure=\frac{2\times Precision\times Recall}{Precision+Recall} \tag{14}$$

4.3　实验对比与结果分析

根据公式(12)–公式(14)的计算方法，分别计算 TFT-Based、TT-Based、FT-Based 和 T-Based 4 种推荐方法的 *Precision*、*Recall* 和 *F-Measure* 值，并用直观图形显示，如图 6–图 8 所示，其中横坐标表示推荐结果靠前(Top$K$)的 K 的不同取值。

从图 6 可知，4 种方法推荐的准确率随着 K 值的增加缓慢提高，最高的是本文的推荐方法。但当 $K\leqslant 15$ 时，4 种推荐方法的准确率都不高，这可能与用户用词的趋同性有关。本来是截然不同的两个物品，在用户没有标注细粒度品类特征的情况下，所描述的属性特征却可能相同，这样会导致与目标用户兴趣不同的物品可能会被推荐，降低商品推荐的准确率。如一件夏

季的蕾丝衬衫和一件秋季的蕾丝外套，都标注"蕾丝、时尚、韩范儿"的标签，被认为是相同或相似度很高的两件衣服会被推荐给目标用户，但欲购买秋装的目标用户可能不喜欢，这也证实了基于文本分析的推荐的弊端。实际上，用户喜欢用图文并茂的形式分享自己喜欢的物品，如果能够识别图片中物品的特征，在研究时将其添加到标签中，或直接提取图片中物品的识别特征，在此基础上进行推荐，将会大大提高基于标签或关键词的推荐准确度。

随着 K 值的增加，各种方法的 *Precision* 都有所提高。提高最快的是本文方法 TFT-Based，不稳定且最慢的是不考虑标签频率权重的方法 T-Based。另外 TFT-Based 相对于单纯考虑标签频率权重的 FT-Based 方法在推荐数目≤15时，准确率相同；当推荐数目>15时，TFT-Based 的准确率比 FT-Based 高。而单纯考虑时间权重的 TT-Based 方法与 T-Based 的准确率相近，却明显低于 TFT-Based 和 FT-Based。这说明，考虑标签频率和时间对用户偏好的影响是必要的，标注频率对用户偏好的影响比标注时间对用户偏好的影响更大。本文数据是按最新优先的顺序获取的，这也说明在没有特殊情况下，用户对某一领域的兴趣偏好在短期内变化不大，需要进一步关注和分析拐点时间对用户兴趣偏好的影响。

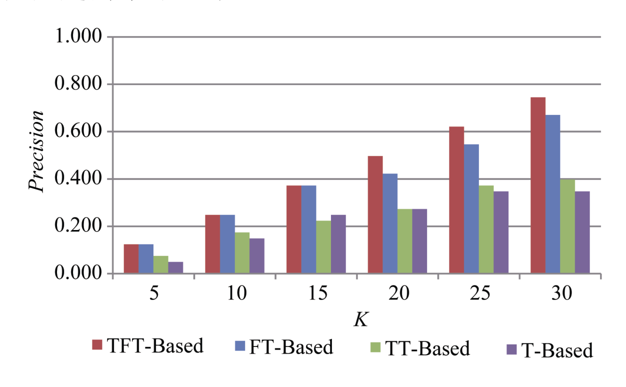


图 6　不同 K 值下的 *Precision* 值比较

从图 7 可知，4 种方法推荐的召回率在 K≤10 时也均较低，最高的也只有不到 40%，这可能与用户标注标签的自由、随意有关。同样一件物品，不同的人看问题的视角不同，兴趣点也不同，所以对同一个特征所用词汇不同，同一件物品所标注的特征也不同。因此，虽然目标用户与推荐用户喜欢并描述了同一个物品，由于标注的标签大相径庭，因而该物品得不到推荐。这给基于标签的商品推荐提出了很大的挑战，如何对同一物品的不同标签实现统一标注，目前还没有很好的方法。结合图片的物品语义特征的分析也许会是一个不错的方案，但如何在体量如此大的标签系统中实现，还需要利用图像技术和大数据处理技术进行尝试。但图 7 中的 4 种方法的召回率随着 K 值的不断增加也呈现上升趋势。而横向来看，本文提出的方法整体上较其他三种方法的召回率都有更好的表现。虽然当 K≤15 时，TFT-Based 与 FT-Based 的召回率相同，变化趋势相同；但当 K>15 时，TFT-Based 的召回率大于 FT-Based，并且差距有不断扩大的趋势。而 FT-Based 的召回率也明显高于 TT-Based 方法，说明标注频率对推荐召回率的影响大于标注时间对推荐召回率的影响，考虑标注频繁度更能提高用户喜欢的物品被推荐的概率。另外，考虑标注时间对用户偏好的影响对推荐结果的召回率作用不明显。在同一 K 值下，TT-Based 与 T-Based 的召回率相差不大，有时 TT-Based 略高，有时 T-Based 略高；在不同 K 值下，这两种方法相差也不大，但总体上，TT-Based 较 T-Based 有更好的表现。

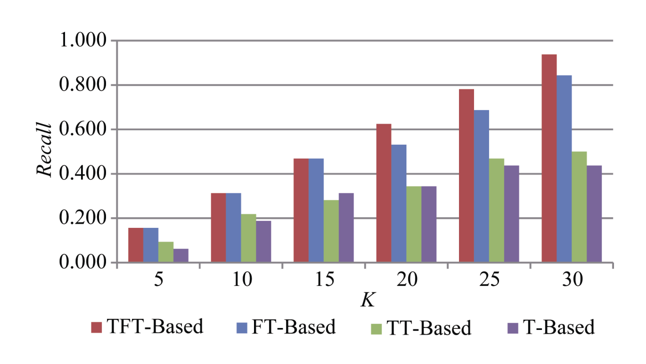


图 7　不同 K 值下的 *Recall* 值比较

F-Measure 值是 *Precision* 和 *Recall* 的综合，从图 8 可看出，随着 K 值的增加，4 种方法的 *F-Measure* 值都呈上升趋势。其中本文方法几乎为线性变化，当 K≤15 时，TFT-Based 与 FT-Based 的 F-Measure 值变化线重合；当 K>15 时，后者的变化曲线低于 TFT-Based，即其 F-Measure 值小于 TFT-Based。而 TT-Based 的 F-Measure 与 T-Based 相近，其变化规律也与准确率和召回率的相似。但 TFT-Based 的 F-Measure 值相对于其他方法总体来看最高，FT-Based 次之，TT-Based 和 T-Based 较小。

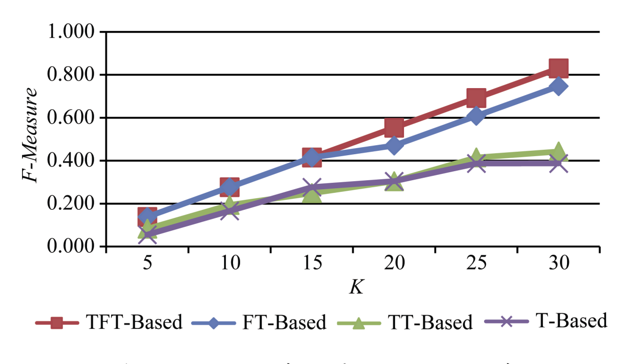


图 8 不同 *K* 值下的 *F-Measure* 值

综合来看，虽然 4 种推荐方法在 *K* 值较小时的准确率、召回率和 *F-Measure* 值都较低，但三种指标值都呈现随 *K* 值增加而上升的趋势，且本文方法的三种指标的表现略高于考虑标注频率对用户偏好影响的推荐方法，这说明用户标注频率对用户偏好的影响显著。但考虑标注时间对用户偏好影响的推荐方法的表现与不考虑标签权重的影响的推荐方法相当，也就是说，标注时间对用户偏好的影响很小，当然，这也许还与获取的标签的时间范围有关。因此，考虑标注时间周期长短以及时间拐点对用户偏好的影响，是需要进一步研究的方向。另外，除了时间因素和标注频率因素，是否还有其他因素(如标注习惯、标签获取方式)的影响，以至于会产生不同的推荐效果，也有待进一步探索。

5 结 语

本文针对现有基于标签的推荐研究中推荐精确度不高的问题，提出一种结合商品标签本体与标签权重的推荐方法。在构建本体时，参照用户标注的标签信息和相关电子商务网站关于商品检索条件，构建基于标签的商品本体。在进行用户偏好建模时，同时考虑用户使用标签的频率与用户兴趣随时间变化两个权重，作为标签对用户重要度权重，也即用户对商品标签的偏好值。之后，计算用户偏好商品标签与商品标签的相似度，用户将获得相似度最高的 K 个商品推荐。实验结果表明，该方法相对于利用标签进行协同过滤推荐方法具有较优的效果，计算的时间和空间的复杂度更小。社会化电子商务中用户自由标注的商品标签不仅可以描述商品特征而且隐含了用户的偏好，但社会化标签在赋予用户自由、自愿管理自己感兴趣的资源权利的同时，也给标签数据的处理带来了巨大的挑战。用户标签用词随意、语义模糊，而整体数量庞大，使得在进行推荐时需要大量工作以规范化标签语义。本文使用商品标签本体来序化标签、优化标签语义，但本体构建本身是一个复杂的工程，还没有通用的、面对动态数据的本体构建方法，这将是进一步研究的方向。

参考文献:

[1] 于洪，李俊华. 结合社交与标签信息的协同过滤推荐算法[J]. 小型微型计算机系统，2013, 34(11): 2467-2471. (Yu Hong, Li Junhua. Collaborative Filtering Recommendation Algorithm Using Social and Tag Information[J]. Journal of Chinese Computer Systems, 2013, 34(11): 2467-2471.)

[2] Ji A T, Yeon C, Kim H, et al. Collaborative Tagging in Recommender Systems [C]// Proceedings of the 20th Australian Joint Conference on Artificial Intelligence. Berlin: Springer-Verlag, 2007: 377-386.

[3] Marinho L B, Schmidt-Thieme L. Collaborative Tag Recommendations [EB/OL]. [2016-05-25]. http://www.springerlink. com/index/m5688r6761448612. pdf.

[4] Nakamoto R, Nakajima S, Miyazaki J, et al. Tag-based Contextual Collaborative Filtering[J]. IAENG International Journal of Computer Science, 2007, 34 (2): 214-219.

[5] Zhao S, Du N, Nauerz A, et al. Improved Recommendation Based on Collaborative Tagging Behaviors[C]//Proceedings of the International Conference on Intelligent User Interfaces. New Mexico: ACM Press, 2008: 413-416.

[6] Gu Y, Yang Z, Kitsuregawa M. Towards Effective Recommendation in Asocial Annotation System Through Group Extraction [EB/OL]. [2011-12-01]. http://db-event. jpn.org /deim2011/proceedings/pdf /f96.pdf.

[7] Niwa S, Doi T, Honiden S. Web Page Recommender System Based on Folksonomy Mining[C]//Proceedings of the 3rd International Conference on Information Technology. 2006: 388-393.

[8] Gemmell J, Shepitsen A, Mobasher B, et al. Personalizing Navigation in Folksonomies Using Hierarchical Tag Clustering [A]// Data Warehousing and Knowledge Discovery [M]. Springer, Berlin, Heidelberg, 2008: 196-205.

[9] 杨丹，曹俊. 基于 Web2.0 的社会性标签推荐系统[J].重庆工学院学报：自然科学版，2008, 22(7): 52-53. (Yang Dan, Cao Jun. Web Page Recommender System Based on Social Tags in Web 2.0 [J]. Journal of Chongqing Institute of Technology, 2008, 22(7): 52-53.)

[10] Zhang Z, Zhou T, Zhang Y. Personalized Recommendation

via Integrated Diffusion on User-Item-Tag Tripartite Graphs [J]. Physica A: Statistical Mechanics and Its Applications, 2010, 389 (1): 179-186.

[11] Li D, Xu Z, Yang M, et al. Item Recommendation in Social Tagging Systems Using Tag Network [J]. Journal of Information and Computational Science, 2013, 10(13): 4057-4066.

[12] Hotho A, Jäschke R, Schmitz C, et al. FolkRank: A Ranking Algorithm for Folksonomies[C]// Proceedings of the 2006 Lernen-Wissensentdeckung-Adaptivität (LWA 2006). 2006: 111-114.

[13] Schmitz C, Hotho A, Jäschke R, et al. Mining Association Rules in Folksonomies[A]// Data Science and Classification [M]. Berlin: Springer-Verlag, 2006: 261-270.

[14] 曹高辉, 毛进. 基于协同标注的 B2C 电子商务个性化推荐系统研究[J]. 图书情报工作, 2008, 52(12): 126-128. (Cao Gaohui, Mao Jin. Research on a Collaborative Tagging System for Personalized Recommendation in B2C Electronic Commerce[J]. Library and Information Service, 2008, 52(12): 126-128.)

[15] 田莹颖. 基于社会化标签系统的个性化信息推荐探讨[J]. 图书情报工作, 2010, 54 (1): 50-54. (Tian Yingying. On Personalized Information Recommendation Based on Social Tagging System[J]. Library and Information Service, 2010, 54 (1): 50-54.)

[16] 邓双义. 基于语义的标签推荐系统关键问题研究[D].上海: 华东师范大学, 2009. (Deng Shuangyi. Research on Key Problems of Tag Recommendation System Based on Semantic [D]. Shanghai: East China Normal University, 2009.)

[17] Rafailidis D, Daras P. The TFC Model: Tensor Factorization and Tag Clustering for Item Recommendation in Social Tagging Systems[J]. IEEE Transactions on Systems, Man, and Cybernetics: System, 2013, 43(3): 673-688.

[18] 郭娣, 赵海燕. 融合标签流行度和时间权重的矩阵分解推荐算法[J]. 小型微型计算机系统, 2016, 37(2): 293-297. (Guo Di, Zhao Haiyan. Matrix Factorization Recommendation Algorithm Fusing Tag Popularity and Time Weight[J]. Journal of Chinese Computer Systems, 2016, 37(2): 293-297.)

[19] Durão F A, Dolog P. Analysis of Tag-Based Recommendation Performance for a Semantic Wiki[C]// Proceedings of the 6th European Semantic Web Conference, Hersonissos, Greece. 2009.

[20] Cheng Y, Qiu G, Bu J J, et al. Model Bloggers' Interests Based on Forgetting Mechanism[C]//Proceedings of the 17th International Conference on World Wide Web. New York: ACM Press, 2008: 1129-1130.

[21] 蔡强, 韩东梅, 李海生, 等. 基于标签和协同过滤的个性化资源推荐[J]. 计算机科学, 2014, 41(1): 69-71, 110. (Cai Qiang, Han Dongmei, Li Haisheng, et al. Personalized Resource Recommendation Based on Tags and Collaborative Filtering[J]. Computer Science, 2014, 41(1): 69-71, 110.)

[22] 赵艳, 王亚民. P2P 环境下基于社会化标签的个性化推荐模型研究[J]. 现代图书情报技术, 2014(5): 50-57. (Zhao Yan, Wang Yamin. Model for Personalized Recommendation Based on Social Tagging in P2P Environment[J]. New Technology of Library and Information Service, 2014(5): 50-57.)

作者贡献声明:

涂海丽: 数据采集, 进行实验, 结果分析, 论文起草;
唐晓波: 提出研究思路, 设计研究方案及框架, 论文最终版本修订。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: 69417380@qq.com。
[1] 涂海丽.tag.txt. 用户、商品及其标签数据。

收稿日期: 2016-12-07
收修改稿日期: 2017-05-04

Building Product Recommendation Model Based on Tags

Tu Haili[1] Tang Xiaobo[2]
[1](School of Economics and Management, East China University of Technology, Nanchang 330013, China)
[2](School of Information Management, Wuhan University, Wuhan 430072, China)

Abstract: [**Objective**] This paper proposes a personalized product recommendation model based on tags in the social e-commerce environment. [**Methods**] First, we calculated users' interests and preferences with the help of tagging frequency and time. Then, we constructed a product ontology of the commercial community based on the tag features and searching conditions of the e-commerce website. Third, we used the ontology to standardize tag semantics, and to classify goods. Fourth, we found clusters containing user preferences, and calculated the similarity between their tags of goods and user preference in the cluster. Finally, we identified the goods which were not tagged but preferred by a specific user. [**Results**] We examined the model with information of 200 randomly selected active users of popular items from the website of FanDongXi. [**Limitations**] Only used the frequency and time factor of the users' tags to calculate their interests and preferences. [**Conclusions**] The proposed method has better performance than the collaborative filtering recommendation based methods.

Keywords: User Tag Product Ontology User Preference Recommendation Model

Springer Nature 开创同行评审者慈善激励制度

得益于 Springer Nature 旗下 *Environmental Earth Sciences* 杂志和非营利人道主义组织"Filter of Hope"的一项合作，同行评审人员开始帮助发展中国家的居民获得安全的饮用水。自 2017 年初该计划实施以来，已在利比里亚、尼加拉瓜、海地、洪都拉斯、俄罗斯、古巴和印度分发了近 600 个生活用水过滤器。该计划通过非营利性合作伙伴关系，首次对同行评议人员在科学出版业中所做的基础性贡献进行奖励。

当评审人员完成 *Environmental Earth Sciences* 杂志的同行评审时，Springer Nature 将在稿件提交系统中进行跟踪，以便对"Filter of Hope"进行相应的捐赠。评审人员还可以选择是否希望期刊在年底特刊中对其所做的评审工作进行答谢。

"Filter of Hope——清洁生活用水"是一个非营利性组织，为 40 多个国家的人们提供服务。他们的目标是通过分发高效的、经济实惠的生活用水过滤器来改变世界。生活用水过滤器能从污染的水源中去除细菌、原生动物和微生物，使其完全达到安全饮用标准。"Filter of Hope"的工作取决于全球各地的分销机构和资助者，包括全球各地的基金会、企业、慈善家庭、学校、教会、人道主义团体和青年人。

"Filter of Hope"创始人 Bart Smelley 表示："感谢 Springer，世界各地的人们现在都可以使用干净的饮用水了。Springer 和我们之间的这种伙伴关系正在改变世界。"

Environmental Earth Sciences 杂志高级编辑 Annett Buettner 说："每一份同行评议都是非常重要的！这是我们开始实施这个计划时想传达出的信息。审稿人是确保出版物的科学诚信和准确性的基础。无数的调研和市场研究表明，同行评审人员不希望期刊对其进行货币激励。这个计划允许我们以小的姿态来答谢审稿人，同时对发展中国家的家庭产生有益的影响。希望其他期刊也能考虑这种合作模式。"

环境地球科学是一个关心人类、自然资源、生态系统、特殊气候或独特地理区域，与地球之间相互作用的，国际性的、多学科的一本期刊。其目的是改善和修复地球的环境，使地球成为生命栖息地。

(编译自：http://www.springer.com/gp/about-springer/media/press-releases/corporate/springer-nature-pioneers-charitable-incentive-system-for-peer-reviewers/15035940)

(本刊讯)